Ensemble Kalman filtering without the intrinsic need for inflation

Marc Bocquet

Nonlin. Processes Geophys., 18, 735–750, 2011
www.nonlin-processes-geophys.net/18/735/2011/
doi:10.5194/npg-18-735-2011

Nonlinear Processes in Geophysics

Ensemble Kalman filtering without the intrinsic need for inflation

M. Bocquet[1,2]

[1]Université Paris-Est, CEREA Joint Laboratory École des Ponts ParisTech/EDF R&D, France
[2]INRIA, Paris Rocquencourt Research Center, France

Received: 8 August 2011 – Revised: 14 October 2011 – Accepted: 16 October 2011 – Published: 20 October 2011

Abstract. The main *intrinsic* source of error in the ensemble Kalman filter (EnKF) is sampling error. External sources of error, such as model error or deviations from Gaussianity, depend on the dynamical properties of the model. Sampling errors can lead to instability of the filter which, as a consequence, often requires inflation and localization. The goal of this article is to derive an ensemble Kalman filter which is less sensitive to sampling errors. A prior probability density function conditional on the forecast ensemble is derived using Bayesian principles. Even though this prior is built upon the assumption that the ensemble is Gaussian-distributed, it is different from the Gaussian probability density function defined by the empirical mean and the empirical error covariance matrix of the ensemble, which is implicitly used in traditional EnKFs. This new prior generates a new class of ensemble Kalman filters, called finite-size ensemble Kalman filter (EnKF-N). One deterministic variant, the finite-size ensemble transform Kalman filter (ETKF-N), is derived. It is tested on the Lorenz '63 and Lorenz '95 models. In this context, ETKF-N is shown to be stable without inflation for ensemble size greater than the model unstable subspace dimension, at the same numerical cost as the ensemble transform Kalman filter (ETKF). One variant of ETKF-N seems to systematically outperform the ETKF with optimally tuned inflation. However it is shown that ETKF-N does not account for all sampling errors, and necessitates localization like any EnKF, whenever the ensemble size is too small. In order to explore the need for inflation in this small ensemble size regime, a local version of the new class of filters is defined (LETKF-N) and tested on the Lorenz '95 toy model. Whatever the size of the ensemble, the filter is stable. Its performance without inflation is slightly inferior to that of LETKF with optimally tuned inflation for small interval between updates, and superior to LETKF with optimally tuned inflation for large time interval between updates.

Correspondence to: M. Bocquet
(bocquet@cerea.enpc.fr)

1 Introduction

The ensemble Kalman filter (EnKF) has become a very popular potential substitute to variational data assimilation in high dimension, because it does not require the adjoint of the evolution model, because of a low storage requirement, because of its natural probabilistic formulation, and because it easily lends itself to parallel computing (Evensen, 2009 and reference therein).

1.1 Errors in the ensemble Kalman filter schemes

The EnKF schemes can be affected by errors of different nature. A flaw of the original scheme (Evensen, 1994) which incompletely took into account the impact of the uncertainty of the observations in the analysis, was corrected by Burgers et al. (1998). They introduced a stochastic EnKF, by perturbing the observations for each member of the ensemble, in accordance with the assumed observational noise. Alternatives to the stochastic schemes are the deterministic ensemble Kalman filters introduced by Anderson (2001); Whitaker and Hamill (2002); Tippett et al. (2003).

Even with this correction, EnKF is known to often suffer from undersampling issues, because it is based on the initial claim that the few tens of members of the ensemble may suffice to represent the first and second-order statistics of errors of a large geophysical system. This issue was diagnosed very early by Houtekamer and Mitchell (1998); Whitaker and Hamill (2002). Indeed, the failure to properly sample leads to an underestimation of the error variances, and ultimately to a divergence of the filter.

Adding to the error amplitude mismatch, undersampling generates spurious correlations, especially at long distance separation as addressed by Houtekamer and Mitchell (1998); Hamill et al. (2001).

The sampling errors are intrinsic deficiencies of the EnKF algorithms. In addition to these, one should also account for model errors that are of external nature for an EnKF scheme. Indeed, they are not due to a flaw in the data assimilation algorithm but to deficiencies in the evolution model.

1.2 Strategies to reduce error

Besides the early correction of the stochastic filter, techniques were devised to correct, or make up for the sampling issues. For both deterministic and stochastic filters, the error amplitude problem can be fixed by the use of an inflation of the ensemble: the anomalies (deviations of the members from the ensemble mean) are scaled up by a factor that accounts for the underestimation of the variances (Anderson and Anderson, 1999; Hamill et al., 2001). Alternatively, the inflation can be additive via stochastic perturbations of the ensemble members, as shown by Mitchell and Houtekamer (1999); Corazza et al. (2002) where it was used to account for the misrepresented model error.

As far as stochastic filters are concerned, Houtekamer and Mitchell (1998, 2001) proposed to use a multi-ensemble configuration, where the ensemble is split into several sub-ensembles. The Kalman gain of one sub-ensemble can be computed from the rest of the ensemble, avoiding the so-called *inbreeding* effect. Remarkably, in a perfect model context, the scheme was shown to avoid the intrinsic need for inflation (Mitchell and Houtekamer, 2009).

Unfortunately, inflation or multi-ensemble configuration do not entirely solve the sampling problem and especially the long-range spurious correlations. These can be addressed in two ways under the name of localization. The first route consists in increasing the rank of the forecast error covariance matrix by applying a Schur product with a short-range admissible correlation matrix (Houtekamer and Mitchell, 2001; Hamill et al., 2001). The second route consists in making the analysis local by assimilating a subset of nearby observations (Ott et al., 2004 and references therein). Though vaguely connected, the two approaches still require common grounds to be understood (Sakov and Bertino, 2010). But alternative methods have emerged, either based on cross-validation (Anderson, 2007a), on multiscale analysis (Zhou et al., 2006), or on empirical considerations (Bishop and Hodyss, 2007).

Many of these techniques introduce additional parameters, such as the inflation factor, the number of sub-ensembles, or the localization length. A few of these parameters can even be made local. They can be chosen from experience gathered on a particular system, or they can be estimated online.

The online estimation methods are adaptive techniques, which is a growing subject. Focussing on the inflation issue, they are based on a specific maximum likelihood estimator of the inflation scaling, or of several scalars that parameterize the error covariance matrices (Mitchell and Houtekamer, 1999; Anderson, 2007b; Brankart et al., 2010) essentially following the ideas of Dee (1995). Another adaptive approach (Li et al., 2009) use the diagnostics of Desroziers et al. (2005).

1.3 Towards objective identification of errors

More straightforward approaches have recently been explored through the identification of the sampling errors. Mallat et al. (1998); Furrer and Bengtsson (2007); Raynaud et al. (2009) put forward a quantitative argument that shows the shortcomings of sampling. Let us define an ensemble of N state vectors $\boldsymbol{x}_k$ in $\mathbb{R}^M$, for $k = 1, \ldots, N$. The empirical mean of the ensemble is

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k \,, \tag{1}$$

and the empirical background error covariance matrix of the ensemble is

$$\mathbf{P} = \frac{1}{N-1}\sum_{k=1}^{N} (\boldsymbol{x}_k - \overline{\boldsymbol{x}})\,(\boldsymbol{x}_k - \overline{\boldsymbol{x}})^{\mathrm{T}} \,. \tag{2}$$

They assume that the ensemble members are drawn from a multivariate Gaussian distribution of unknown covariance matrix $\mathbf{B}$, that generally differs from the empirically estimated covariance matrix $\mathbf{P}$. Then the variance of each entry of $\mathbf{P}$ can be assessed using Wick's theorem (Wick, 1950):

$$\mathrm{E}\left([\mathbf{P}-\mathbf{B}]_{ij}^2\right) = \frac{1}{N}\left([\mathbf{B}]_{ii}[\mathbf{B}]_{jj} + [\mathbf{B}]_{ij}^2\right), \tag{3}$$

with $i, j = 1, \ldots, M$ indexing the state space grid-cells; E is the expectation operator of the Gaussian process and $[\mathbf{C}]_{ij}$ generically denotes entry (i, j) of matrix $\mathbf{C}$.

In particular one obtains the average of the error on the estimated variances in $\mathbf{B}$

$$\mathrm{E}\left([\mathbf{P}-\mathbf{B}]_{ii}^2\right) = \frac{2}{N}[\mathbf{B}]_{ii}^2, \tag{4}$$

which has been used in an ensemble of assimilations (Raynaud et al., 2009). Considering covariances at long distance ($i \neq j$), $[\mathbf{B}]_{ij}$ is expected to vanish for most geophysical systems. And yet the errors in estimating $[\mathbf{B}]_{ij}$

$$\mathrm{E}\left([\mathbf{P}-\mathbf{B}]_{ij}^2\right) \sim \frac{1}{N}[\mathbf{B}]_{ii}[\mathbf{B}]_{jj}, \tag{5}$$

are all but vanishing for a small ensemble. The impact of these errors on the analysis can be objectively estimated using the results of van Leeuwen (1999); Furrer and Bengtsson (2007); Sacher and Bartello (2008).

This type of approach may offer objective solutions to account for sampling errors. However incorporating them into data assimilation scheme is not straightforward. For instance, the objective identification of the covariance errors Eq. (3) depends on the true covariances, which are unknown, and some approximate closure is needed.

The Gaussian assumption made by these authors on the distribution from which the ensemble is generated should be regarded as an approximation in the context of ensemble Kalman filtering since such an ensemble often results from

the propagation by a possibly nonlinear dynamical model. However this assumption allows to perform analytical computation using the properties of Gaussian distributions. Besides if the analysis of the data assimilation system only requires first- and second-order moments, higher-order moments are irrelevant for the update, although certainly not for the global performance of a filter. Following these authors, we shall use this statistical assumption.

1.4 Objectives and outline

In the context of ensemble Kalman filtering, the first objective of this article is to build a prior, to be used in the analysis step. Working on the first- and second-order empirical moments of the ensemble, a traditional ensemble Kalman filter performs an update as if the prior distribution was given by a Gaussian defined by the empirical moments $\overline{\boldsymbol{x}}$ and $\mathbf{P}$. Instead, our prior of the true state is conditioned on the entire forecast ensemble, not only its first- and second-order empirical moments. Knowing about the discrete nature of the ensemble, it should partly or completely account for the sampling flaws.

Our goal is, within the framework of ensemble Kalman filtering, to perform a Bayesian analysis with this new prior. In Sect. 2, such a prior is derived.

The use of this prior in the analysis will result in the definition of a new class of algorithms for high-dimensional filtering that are exploited in Sect. 3, the finite-size (i.e. finite-sample) ensemble Kalman filters (denoted EnKF-N). We shall study one of its variant, which is an extension of the ensemble transform Kalman filter (ETKF) of Hunt et al. (2007), that we call the finite-size ensemble transform Kalman filter (ETKF-N).

In Sect. 4, the new filters are applied to the Lorenz '63 and Lorenz '95 models. Their performance is compared to ETKF. The new filters do not seem to require inflation. Unfortunately, like any ensemble Kalman filter, ETKF-N diverges for small ensemble sizes in the Lorenz '95 case. It does require localization. This shows that the new filters do not entirely solve the sampling issue, and the reason for this is discussed in Sect. 5. Yet, a local variant of ETKF-N, the finite-size local ensemble transform Kalman filter (LETKF-N), can be built. It is tested on the Lorenz '95 toy model, and compared to the local ensemble transform Kalman filter (LETKF). The main goal of introducing LETKF-N is to examine whether the need for inflation is still avoided, in spite of the imbalance that localization is known to generate. In Sect. 6, the results are summarized. A few leads to go further are also discussed.

In this article, model error is not considered. It is assumed throughout this study that the model is perfect. Therefore, in this study, inflation is meant to compensate for sampling errors (hence the adjective *intrinsic* in the title). Theoretically, (additive or multiplicative) inflation for model error is a rather distinct subject from inflation for sampling errors, even though it is difficult to untangle the two in practice.

The filters derived in this article should be applicable to (very) high-dimensional geophysical systems. This requires that only a small ensemble can be propagated between updates (typically no more than 100 members).

2 Accounting for sampling errors

We would like to reformulate the traditional analysis step of the EnKF. The prior (or previous forecast) is the focus of the reasoning. The prior that is usually used in the EnKF is given by the prior pdf of the state vector $\boldsymbol{x}$, a vector in $\mathbb{R}^M$, conditional on the empirical mean $\overline{\boldsymbol{x}}$ and on the empirical background error covariance matrix $\mathbf{P}$, defined in Eqs. (1) and (2). Moreover this conditional pdf of the prior $p(\boldsymbol{x}|\overline{\boldsymbol{x}},\mathbf{P})$ is implicitly assumed to be Gaussian. Lacking further information, it is the more natural distribution knowing its first- and second-order moments.

2.1 Getting more from the ensemble

Unfortunately, information is lost: this prior does not take into account the fact that $\overline{\boldsymbol{x}}$ and $\mathbf{P}$ originate from sampling. That is why we aim at computing the prior pdf of $\boldsymbol{x}$ conditional on the ensemble, $p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$. It is assumed that the members of the ensemble are independently drawn from a multivariate Gaussian distribution of mean state $\boldsymbol{x}_b$ and covariance matrix $\mathbf{B}$. As argued in the introduction this assumption leads to an approximation, since the ensemble members are rather samples of a (more or less) non-Gaussian distribution (Bocquet et al., 2010; Lei et al., 2010). There is no point in modelling higher-order moments of the statistics prior to the analysis, since the analysis of the Kalman filter only uses the first- and second-order moments. The moments $\boldsymbol{x}_b$ and $\mathbf{B}$ of the true sampled distribution are unknown a priori and may differ from $\overline{\boldsymbol{x}}$ and $\mathbf{P}$.

Summing over all potential $\boldsymbol{x}_b$ and $\mathbf{B}$, where $\mathbf{B}$ is a positive definite matrix, the prior pdf reads

$$\begin{aligned} &p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\\ &=\int \mathrm{d}\boldsymbol{x}_b\mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{x}_b,\mathbf{B})\,p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\\ &=\int \mathrm{d}\boldsymbol{x}_b\mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})\,p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N). \end{aligned} \tag{6}$$

The symbol $\mathrm{d}\mathbf{B}$ corresponds to the Lebesgue measure on all independent entries $\prod_{i\leq j}^{M}\mathrm{d}[\mathbf{B}]_{ij}$, but the integration is restricted to the cone of positive definite matrices. From the first to the second line, we used the fact that under the assumption of Gaussianity of the prior pdf of the errors, the conditioning of $p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{x}_b,\mathbf{B})$ on the ensemble is redundant, since the pdf is completely characterized by $\boldsymbol{x}_b$, and $\mathbf{B}$. Bayes' rule can be applied to $p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$, so that

$$p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \frac{1}{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)} \times \int \mathrm{d}\boldsymbol{x}_b \mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})\, p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N|\boldsymbol{x}_b,\mathbf{B})\, p(\boldsymbol{x}_b,\mathbf{B}). \quad (7)$$

The probability densities that are conditional on $\boldsymbol{x}_b$ and $\mathbf{B}$ can be written explicitly thanks to the Gaussian assumptions. The first one in Eq. (7) would be the prior of $\boldsymbol{x}$, if one knew the exact mean and error covariance matrix. The second one is the likelihood of the members to be drawn from the Gaussian distribution of the same mean and error covariance matrix (similarly to Dee, 1995). The third pdf in the integral of Eq. (7) is a prior on the background statistics (an *hyperprior*) whose choice will be discussed later. Writing explicitly the two Gaussian pdfs in the integral of Eq. (7) and re-organizing the terms, one gets

$$p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \propto \int \mathrm{d}\boldsymbol{x}_b \mathrm{d}\mathbf{B} \exp(-\mathcal{L}(\boldsymbol{x},\boldsymbol{x}_b,\mathbf{B}))\, p(\boldsymbol{x}_b,\mathbf{B}), \quad (8)$$

where

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{x}_b,\mathbf{B}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b) + \frac{1}{2}(N+1)\ln|\mathbf{B}| + \frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{x}_k-\boldsymbol{x}_b), \quad (9)$$

where $|\mathbf{B}|$ denotes the determinant of $\mathbf{B}$.

2.2 Choosing priors for the background statistics

For the filters designed in this article, like for any (very) high-dimensional ensemble-based Kalman filters, information on the background error statistics can only be transported by the ensemble between analyzes. Passing along information on the full statistics of the errors requires too much storage. That is one reason why the EnKF was preferred over the impractical extended Kalman filter. Still, we have to make a priori assumptions on (the statistics of) $\boldsymbol{x}_b$ and $\mathbf{B}$.

The most popular one in multivariate statistics is Jeffreys' prior. It maximizes the information that will be gained in any subsequent analysis made with that prior (making it as much less informative as possible). It is known that Jeffrey's prior for the couple $(\boldsymbol{x}_b,\mathbf{B})$ is not satisfying in practice, and one should make the independence assumption (Jeffreys, 1961):

$$p(\boldsymbol{x}_b,\mathbf{B}) \equiv p_{\mathrm{J}}(\boldsymbol{x}_b,\mathbf{B}) = p_{\mathrm{J}}(\boldsymbol{x}_b)\, p_{\mathrm{J}}(\mathbf{B}) \quad (10)$$

and compute the Jeffreys' priors for $\boldsymbol{x}_b$ and $\mathbf{B}$ separately. Jeffreys' choice corresponds to

$$p_{\mathrm{J}}(\boldsymbol{x}_b) = 1, \quad p_{\mathrm{J}}(\mathbf{B}) = |\mathbf{B}|^{-\frac{M+1}{2}}, \quad (11)$$

where M is the dimension of the state space. The fact that $p_{\mathrm{J}}(\boldsymbol{x}_b,\mathbf{B})$ cannot be normalized is not truly an issue, like for any non-informative priors in Bayesian statistics, provided (as far as we are concerned) that integral Eq. (7) is proper. The prior of $\mathbf{B}$ has some important properties that are essential for this study. First, it is invariant by any reparameterization of state vectors. Consider the change of state variables $\boldsymbol{x} = \mathbf{F}\boldsymbol{x}'$, where $\mathbf{F}$ is a non-singular matrix in state space. It translates to $\mathbf{B} = \mathbf{F}\mathbf{B}'\mathbf{F}^{\mathrm{T}}$ for the error covariance matrices. The Jacobian of this change of variables for $\mathbf{B}$ is (see for instance Muirhead, 1982)

$$\mathrm{d}\mathbf{B} = |\mathbf{F}|^{M+1}\mathrm{d}\mathbf{B}', \quad (12)$$

so that

$$p_{\mathrm{J}}(\mathbf{B})\mathrm{d}\mathbf{B} = |\mathbf{F}\mathbf{B}'\mathbf{F}^{\mathrm{T}}|^{-\frac{M+1}{2}}|\mathbf{F}|^{M+1}\mathrm{d}\mathbf{B}' = p_{\mathrm{J}}(\mathbf{B}')\mathrm{d}\mathbf{B}'. \quad (13)$$

This justifies the power $(M+1)/2$. Besides we want the hyperprior to lead to asymptotic Gaussianity: in the limit of a large ensemble, this choice should lead to the usual Gaussian prior used in EnKF analysis. This will be checked in Sect. 3.

2.3 Effective $\mathcal{J}_b$ functional

Choosing the prior $p(\boldsymbol{x}_b,\mathbf{B}) \equiv p_{\mathrm{J}}(\boldsymbol{x}_b)\, p_{\mathrm{J}}(\mathbf{B})$, the integration on $\boldsymbol{x}_b$ in Eq. (8) is straightforward and leads to

$$p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \propto \int \mathrm{d}\mathbf{B} \exp(-\mathcal{J}(\boldsymbol{x},\mathbf{B})), \quad (14)$$

where

$$\mathcal{J}(\boldsymbol{x},\mathbf{B}) = \frac{1}{2}\frac{N}{N+1}(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}}) + \frac{N+M+1}{2}\ln|\mathbf{B}| + \frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k-\overline{\boldsymbol{x}})^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{x}_k-\overline{\boldsymbol{x}}). \quad (15)$$

This functional can be compactly written as

$$\mathcal{J}(\boldsymbol{x},\mathbf{B}) = \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}\mathbf{B}^{-1}\right) + \frac{N+M+1}{2}\ln|\mathbf{B}|, \quad (16)$$

where

$$\mathbf{A} = \frac{N}{N+1}(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathrm{T}} + (N-1)\mathbf{P}. \quad (17)$$

Like for most ensemble Kalman filters, especially ensemble transform Kalman filters, it is assumed in the following that $\boldsymbol{x}-\overline{\boldsymbol{x}}$ belongs to the vector space $\mathcal{V}$ spanned by the anomalies $\boldsymbol{x}_k-\overline{\boldsymbol{x}}$. Because in the context of high-dimensional Kalman filtering $\mathbf{A}$ is rank-deficient

$$\mathrm{rank}(\mathbf{A}) \leq N-1 \ll M, \quad (18)$$

integral Eq. (14) turns out to be improper. The problem can be circumvented. Indeed, the $\mathbf{B}$ matrices to integrate on are merely test positive-definite matrices representing potential error covariance matrices. We could choose to integrate on a relevant subspace rather than on all positive definite matrices. We are merely interested in the matrices that act on $\mathcal{V}$ only, because the state vector lies in $\overline{\boldsymbol{x}}+\mathcal{V}$. Integration on the other

matrices will produce an infinite volume factor with no dependence on $\boldsymbol{x}$, that can be subtracted from the final effective functional. On more rigorous grounds, one can extend matrix $\mathbf{A}$ to a full rank positive definite matrix $\mathbf{A}_\epsilon = \mathbf{A} + \epsilon \mathbf{I}_M$, where $\mathbf{I}_M$ is the identity matrix of state space, and $\epsilon > 0$. Then the integral in Eq. (14) becomes proper. After the integration, one can let ϵ goes to 0. A diverging term depending on ϵ only, and hence of no interest, can then be safely ignored.

To perform the integration on $\mathbf{B}$ in Eq. (14), one can proceed to the change of variables $\mathbf{B} = \mathbf{A}_\epsilon^{1/2} \Omega \mathbf{A}_\epsilon^{1/2}$. From Eq. (12), the Jacobian of this change of variable is

$$\mathrm{d}\mathbf{B} = |\mathbf{A}_\epsilon|^{\frac{M+1}{2}} \mathrm{d}\Omega. \tag{19}$$

Therefore, the dependence in $\boldsymbol{x}$ through $\mathbf{A}_\epsilon$ can be extracted from the integral:

$$\begin{aligned} & p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \\ & \propto |\mathbf{A}_\epsilon|^{-N/2} \int \mathrm{d}\Omega |\Omega|^{-(N+M+1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\Omega^{-1}\right) \\ & \propto |\mathbf{A}_\epsilon|^{-N/2} \propto |\mathbf{A}|^{-N/2}. \end{aligned} \tag{20}$$

It is important to realize that the last determinant of $\mathbf{A}$ actually applies to the restriction of the linear operator represented by $\mathbf{A}$ in the canonical basis of subspace $\mathcal{V}$, which is of dimension lower or equal to $N-1$, and is, by this definition, not singular.

From the expression of $p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$, we deduce the background functional to be used in the subsequent analysis of our variant of the EnKF:

$$\begin{aligned} \mathcal{J}_b(\boldsymbol{x}) &= -\ln p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \frac{N}{2}\ln|\mathbf{A}| + \mathrm{Cst} \\ &= \frac{N}{2}\ln\left|\frac{N}{N+1}(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathrm{T}} + (N-1)\mathbf{P}\right|, \end{aligned} \tag{21}$$

up to some irrelevant constant. Let us remark that the mean of the ensemble $\overline{\boldsymbol{x}}$ is the mean and mode of $p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$.

2.4 Alternate $\mathcal{J}_b$ functional

One can argue against the choice of $p_J(\boldsymbol{x}_b) = 1$. It might be considered too weakly informative. However as an hyperprior, it provides information before the observation, but also before exploiting the ensemble. So, whatever information is passed on to the subsequent analysis, it is weak, unless the information content of the ensemble is weak and the observation are not dense (small ensemble size, sparse/infrequent observation).

One alternative to the uniform distribution is to use a climatology for $\boldsymbol{x}_b$. It is not tested in this study. However it was recently demonstrated in the context of ensemble Kalman filtering that such an approach is helpful for sparsely observed systems (Gottwald et al., 2011). Another alternative is to make specific choices for $\boldsymbol{x}_b$. Equation (6)

$$p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = \int \mathrm{d}\boldsymbol{x}_b \mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B}) p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N), \tag{22}$$

would be affected in the following way. The probability density $p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})$ is conditional on the knowledge of $\mathbf{B}$ and $\boldsymbol{x}_b$. For this density, we additionally assume a great confidence in $\overline{\boldsymbol{x}}$, like any standard EnKF, so that the first guess $\boldsymbol{x}_b$ of the sampled prior is believed to be very close to $\overline{\boldsymbol{x}}$ and $p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B}) \simeq p(\boldsymbol{x}|\overline{\boldsymbol{x}},\mathbf{B})$. This assumption can be wrong for small ensemble size. Therefore:

$$\begin{aligned} & p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \\ & \simeq \int \mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\overline{\boldsymbol{x}},\mathbf{B}) \int \mathrm{d}\boldsymbol{x}_b\, p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N). \end{aligned} \tag{23}$$

The rest of the derivation is fundamentally unchanged. The final background functional reads

$$\mathcal{J}_b^{\mathrm{alt}}(\boldsymbol{x}) = \frac{N}{2}\ln\left|(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathrm{T}} + (N-1)\mathbf{P}\right|. \tag{24}$$

However the disappearance of the $N/(N+1)$ factor is not cosmetic, and may have consequences that are investigated later.

3 Finite-size ensemble transform Kalman filter

Because $\mathcal{J}_b$ and $\mathcal{J}_b^{\mathrm{alt}}$ are not quadratic, it is clear that the analysis should be variational, in a similar flavor as the maximum likelihood ensemble filter (Zupanski, 2005; Carrassi et al., 2009). As such it can accommodate nonlinear observation operators. Therefore, in this study, the analysis step will be variational, similarly to 3D-Var. One should minimize the cost function

$$\mathcal{J}_a(\boldsymbol{x}) = \mathcal{J}_o(\boldsymbol{x}) + \mathcal{J}_b(\boldsymbol{x}), \tag{25}$$

with

$$\mathcal{J}_o(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y}-H(\boldsymbol{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{y}-H(\boldsymbol{x})), \tag{26}$$

where $\boldsymbol{y}$ is the observation vector in observation space $\mathbb{R}^d$, $\mathbf{R}$ is the observation error covariance matrix, and H is the observation operator.

We shall call finite-size (or finite-sample) ensemble Kalman filters (EnKF-N), the ensemble Kalman filters that could be generated using this type of $\mathcal{J}_b$ term in the analysis step of the filter. In the following, the focus will be on the ensemble transform Kalman filter (ETKF) variant, following Hunt et al. (2007). The analysis is expressed as an element of subspace $\overline{\boldsymbol{x}} + \mathcal{V}$. The state vector is characterized by a set of (redundant) coordinates $\{w_k\}_{k=1,\ldots,N}$ in the ensemble space:

$$\boldsymbol{x} = \overline{\boldsymbol{x}} + \sum_{k=1}^{N} w_k (\boldsymbol{x}_k - \overline{\boldsymbol{x}}). \tag{27}$$

If $\mathbf{X}_k = \boldsymbol{x}_k - \overline{\boldsymbol{x}}$ are the anomalies, and $\mathbf{X}$ the matrix of these anomalies, $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_N)$, then $\boldsymbol{x} - \overline{\boldsymbol{x}} = \mathbf{X}\boldsymbol{w}$. Hence, one has

$$\mathbf{A} = \frac{N}{N+1}\mathbf{X}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} + \mathbf{X}\mathbf{X}^{\mathrm{T}}. \tag{28}$$

Recall that $|\mathbf{A}|$ represents the determinant of the linear operator related to $\mathbf{A}$ but restricted to subspace $\mathcal{V}$. In the same subspace, the linear operator related to $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is invertible, of inverse denoted $\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}$. One gets

$$\begin{aligned}|\mathbf{A}| &= \left|\frac{N}{N+1}\mathbf{X}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}+\mathbf{X}\mathbf{X}^{\mathrm{T}}\right| \\ &= \left|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right|\left|\mathbf{I}_{\mathcal{V}}+\frac{N}{N+1}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right| \\ &\propto 1+\frac{N}{N+1}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\boldsymbol{w}. \end{aligned} \tag{29}$$

There is a subtlety that we need to develop on, and which generalizes the clear explanation given by Hunt et al. (2007).

3.1 Gauge-invariance of the parameterization

As a family of vectors, the anomalies are not independent since $\sum_{k=1}^{N}\mathbf{X}_k=\mathbf{0}$. Therefore parameterizing $\mathcal{J}_b(\boldsymbol{x})=\mathcal{J}_b(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})$ with $\boldsymbol{w}$ entails a so-called *gauge invariance* (a redundancy in $\boldsymbol{w}$): $\mathcal{J}_b(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})$ is invariant under a shift of all w_k by a same constant. The number of degrees of freedom of this invariance is given by the dimension of the kernel of $\mathbf{X}$, which is at least one according to the previous remark.

The expression given by Eq. (29) is not invariant under rotations of $\boldsymbol{w}$. We could make it invariant by using the freedom of the gauge invariance. We can fix this gauge by choosing to minimize the cost function over the $\boldsymbol{w}$ that have a null orthogonal projection on the kernel of $\mathbf{X}$:

$$\left(\mathbf{I}_N-\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\right)\boldsymbol{w}=\mathbf{0}. \tag{30}$$

With this constraint, $|\mathbf{A}|$ is proportional to $1+\frac{N}{N+1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$. This is cumbersome to enforce though. Instead, to perform the same task, a gauge-fixing term

$$\mathcal{G}(\boldsymbol{w})=\frac{N}{N+1}\boldsymbol{w}^{\mathrm{T}}\left(\mathbf{I}_N-\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\right)\boldsymbol{w}, \tag{31}$$

is inserted into the cost function Eq. (25)

$$\mathcal{J}_a(\boldsymbol{x})=\mathcal{J}_o(\boldsymbol{x})+\frac{N}{2}\ln(|\mathbf{A}|), \tag{32}$$

yielding an augmented cost function

$$\widetilde{\mathcal{J}}_a(\boldsymbol{w})=\mathcal{J}_o(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})+\frac{N}{2}\ln(|\mathbf{A}|+\mathcal{G}(\boldsymbol{w})). \tag{33}$$

For instance, in the case where $\mathrm{rank}(\mathbf{A})=N-1$, one has

$$\mathcal{G}(\boldsymbol{w})=\frac{1}{N+1}\left(\sum_{k=1}^{N}w_k\right)^2. \tag{34}$$

Because ln is a monotonically increasing function, one gets $\widetilde{\mathcal{J}}_a(\boldsymbol{w})\geq\mathcal{J}_a(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})$, for all $\boldsymbol{w}$ in $\mathbb{R}^N$, with equality if and only if $\mathcal{G}(\boldsymbol{w})=0$. Moreover, for any $\boldsymbol{x}$ there is a $\boldsymbol{w}^\star$ in the kernel of $\mathbf{X}$ ($\mathcal{G}(\boldsymbol{w}^\star)=\mathbf{0}$) such that $\mathcal{J}_a(\boldsymbol{x})=\widetilde{\mathcal{J}}_a(\boldsymbol{w}^\star)$. As a consequence, the two cost functions $\widetilde{\mathcal{J}}_a(\boldsymbol{w})$ and $\mathcal{J}_a(\boldsymbol{x})$ have the same minimum. Note that this implies that at the minimum $\boldsymbol{w}^a$ of $\widetilde{\mathcal{J}}_a$, one has $\mathcal{G}(\boldsymbol{w}^a)=\mathbf{0}$.

Hence, instead of Eq. (25) one can use the cost function with a gauge-fixing term:

$$\begin{aligned}\widetilde{\mathcal{J}}_a(\boldsymbol{w}) &= \frac{1}{2}(\boldsymbol{y}-H(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w}))^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{y}-H(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})) \\ &\quad+\frac{N}{2}\ln\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right). \end{aligned} \tag{35}$$

Cost function Eq. (35) is not necessarily convex because the ln function is concave. Let us assume a linear observation operator, or linearized around the innovation $\boldsymbol{y}-H(\overline{\boldsymbol{x}})$. Then a minimum always exists since for a linear observation operator, $\mathcal{J}_o(\boldsymbol{x})$ is convex in $\boldsymbol{w}$, and

$$\widetilde{\mathcal{J}}_b(\boldsymbol{w})=\frac{N}{2}\ln\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right), \tag{36}$$

is a monotonically increasing function when the norm of $\boldsymbol{w}$ goes to infinity. Conversely, the cost function may have several minima (see Appendix A). As a consequence the nature of the minimizer, as well as the first guess of the iterative optimization, may have an impact on the result. The first guess of the iterative minimization was chosen to be $\boldsymbol{w}=\mathbf{0}$, which favors the prior against the observation if several minima do exist. Even though it may sound wiser to favor observation, the choice $\boldsymbol{w}=\mathbf{0}$ is clearly simpler.

3.2 Posterior ensemble

Once $\boldsymbol{w}^a$ is obtained as the minimizer of Eq. (35), the posterior state estimate is given by

$$\boldsymbol{x}^a=\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w}^a. \tag{37}$$

We wish to compute a local approximation of the error covariances at the minimum. The Hessian of $\widetilde{\mathcal{J}}_b$ can be computed in ensemble space:

$$\widetilde{\mathcal{H}}_b=\nabla^2_{\boldsymbol{w}}\widetilde{\mathcal{J}}_b(\boldsymbol{w})=N\frac{\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right)\mathbf{I}_N-2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}}{\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right)^2}. \tag{38}$$

The Hessian of the observation term is

$$\widetilde{\mathcal{H}}_o=\nabla^2_{\boldsymbol{w}}\mathcal{J}_o(\overline{\boldsymbol{x}}+\mathbf{X}\boldsymbol{w})=(\mathbf{H}\mathbf{X})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{X}, \tag{39}$$

where $\mathbf{H}$ is the tangent linear of H. The analysis error covariance matrix $\widetilde{\mathbf{P}}_a$ in ensemble space is approximately obtained from the inverse matrix of the total Hessian at the minimum

$$\widetilde{\mathbf{P}}_a\simeq\widetilde{\mathcal{H}}_a^{-1}, \tag{40}$$

where $\widetilde{\mathcal{H}}_a=\widetilde{\mathcal{H}}_b(\boldsymbol{w}^a)+\widetilde{\mathcal{H}}_o(\boldsymbol{w}^a)$. Note that $\widetilde{\mathcal{H}}_a$ must be positive definite by construction, even though $\widetilde{\mathcal{H}}_b(\boldsymbol{w}^a)$ is not necessarily so.

Then, a posterior ensemble can be obtained from the square root of $(N-1)\widetilde{\mathbf{P}}_a$. More precisely, the posterior

ensemble anomalies, in ensemble space, are given by the columns $\mathbf{W}_k^a$ of the transform matrix

$$\mathbf{W}^a = \left((N-1)\widetilde{\mathbf{P}}_a\right)^{1/2}\mathbf{U}, \tag{41}$$

where $\mathbf{U}$ is an arbitrary orthogonal matrix that preserves the ensemble mean: $\mathbf{U}\boldsymbol{u} = \boldsymbol{u}$ where $\boldsymbol{u} = (1,\ldots,1)^{\mathrm{T}}$. The degrees of freedom introduced by $\mathbf{U}$ allow to span the ensemble space of any ensemble square root Kalman filter (Sakov and Oke, 2008). Accordingly, the posterior ensemble in state space is given for $k = 1,\ldots,N$ by

$$\boldsymbol{x}_k^a = \boldsymbol{x}^a + \mathbf{X}\mathbf{W}_k^a. \tag{42}$$

Let us check that the posterior ensemble is centered on $\boldsymbol{x}^a$. To do so, one has to verify that $\boldsymbol{u}$ is in the kernel of $\mathbf{X}\mathbf{W}^a$. If we can prove that $\boldsymbol{u}$ is an eigenvector of $\widetilde{\mathbf{P}}_a$, then $\mathbf{X}\mathbf{W}^a\boldsymbol{u} \propto \mathbf{X}\boldsymbol{u} = \mathbf{0}$. The eigenvectors of $\widetilde{\mathbf{P}}_a$ are those of the Hessian $\widetilde{\mathcal{H}}_a$ at the minimum. Since $\mathcal{J}_o(\overline{\boldsymbol{x}} + \mathbf{X}\boldsymbol{w})$ is gauge invariant, it is easy to check that $\boldsymbol{u}$ is in the kernel of the Hessian $\widetilde{\mathcal{H}}_o$. (Note that this remark also applies without approximation to nonlinear observation operators.) As for $\widetilde{\mathcal{J}}_b$ whose gauge-invariance has been intentionally broken, the argument cannot apply. But it was seen earlier that at the minimum $\mathcal{G}(\boldsymbol{w}^a) = 0$. In particular, one has $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{w}^a = 0$. As a consequence, it is clear from Eq. (38) that $\boldsymbol{u}$ is an eigenvector of $\widetilde{\mathcal{H}}_b$, of eigenvalue $N(1+1/N+(\boldsymbol{w}^a)^{\mathrm{T}}\boldsymbol{w}^a)^{-1}$. Therefore the posterior ensemble is centered on $\boldsymbol{x}_a$. This property is important for the consistency and ultimately the stability and performance of the filter (Wang et al., 2004; Livings et al., 2008; Sakov and Oke, 2008).

The new filters are based on several mild approximations that are imposed by the non-Gaussianity of the prior. Firstly, one might not sample the right minimum when there are several of them (see Appendix A). Or the right estimator could be the average rather than a mode of the posterior pdf. Secondly, and unlike the Gaussian case, the inverse Hessian is only a local approximation of the analysis error covariance matrix (Gejadze et al., 2008).

3.3 Asymptotic Gaussianity

When the ensemble size goes to large $N \to \infty$, the ln term in the background part of cost function Eq. (36), must decrease to smaller, yet always positive, values. So should $\varepsilon \equiv \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \sum_{k=1}^N w_k^2$. Therefore, in this limit, one has

$$\begin{aligned}\widetilde{\mathcal{J}}_b &= \frac{N}{2}\ln\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right)\\ &\simeq \frac{1}{2}+\frac{N-1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+O\left(N^{-1},N^{-1}\varepsilon,\varepsilon^2\right),\end{aligned} \tag{43}$$

and the ETKF of Hunt et al. (2007) is recovered (assuming $\mathbf{U}$ is the identity matrix).

3.4 Algorithm

The variant of the finite-size EnKF that has just been described is the finite-size ensemble transform Kalman filter (ETKF-N). The numerical implementation is similar to that of Harlim and Hunt (2007) (see Algorithm 2). The pseudo-code for ETKF-N is:

1. Obtain the forecast ensemble $\{\boldsymbol{x}_k\}_{k=1,\cdots,N}$ from the model propagation of the previous ensemble analysis.
2. Form the mean $\overline{\boldsymbol{x}}$, and the anomaly matrix $\mathbf{X}$, necessary for the evaluation of cost function Eq. (35).
3. Minimize cost function Eq. (35) iteratively starting with $\boldsymbol{w} = \mathbf{0}$, to obtain $\boldsymbol{w}^a$.
4. Compute $\boldsymbol{x}^a$ and the Hessian $\widetilde{\mathcal{H}}_a$, from Eq. (37), Eq. (38), and Eq. (39).
5. Compute $\mathbf{W}^a = \left(\widetilde{\mathcal{H}}_a/(N-1)\right)^{-1/2}\mathbf{U}$.
6. Generate the new ensemble: $\boldsymbol{x}_k^a = \boldsymbol{x}^a + \mathbf{X}\mathbf{W}_k^a$.

The complexity is the same as that of ETKF. The minimization of the analysis cost function, which is already well conditioned by construction, might be longer in such non-quadratic, and even non-convex context. However, the minimization remains in ensemble space, and is almost negligible in cost for high-dimensional applications with an ensemble size in the range of 10–100.

3.5 Interpretation

The influence of the background term of the cost function, $\widetilde{\mathcal{J}}_b = \frac{N}{2}\ln\left(1+1/N+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right)$, within the full cost function Eq. (35), is compared to its counterpart in ETKF, $\widetilde{\mathcal{J}}_b = \frac{N-1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$. Firstly, let us assume that the innovation is such that, in the ETKF system, the analysis is driven away from the ensemble mean:

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \sum_{k=1}^N w_k^2 \geq O(1). \tag{44}$$

In the ETKF-N system, the constraint enforced by the background term would be alleviated by the presence of the ln function. Therefore, in the same situation (same innovation), ETKF-N would be more controlled by the observation than ETKF. In particular, larger deviations from the ensemble mean would be allowed. It is reminiscent of the way the Huber norm operates (Huber, 1973).

Secondly, assume that the innovation drives the ETKF system towards an analysis close to the ensemble mean

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \sum_{k=1}^N w_k^2 \ll 1. \tag{45}$$

From Eq. (43), it is clear that the ETKF-N system is in a similar regime. However, because of the $1/N$ offset in the ln

function, the prior term cannot vanish even when the ensemble mean is taken as the optimal state. This is confirmed by the inverse of the Hessian $\widetilde{\mathcal{H}}_b$, the contribution of the prior to $\widetilde{\mathbf{P}}_a$, which is $N^{-1}(1+1/N)$ at $\boldsymbol{w}^a = \mathbf{0}$, instead of N^{-1}. This also corresponds to the residual $1/2$ term in $\widetilde{\mathcal{J}}_b$ of Eq. (43). Algebraically, this offset comes in the formula by the integration on $\boldsymbol{x}_b$: this *blurring* tells the system not to trust the ensemble mean entirely at finite N.

We believe this is the same term $1+1/N$ that was diagnosed by Sacher and Bartello (2008), who showed that, for a Gaussian process, the dispersion of the ensemble around the mean of the Gaussian should be $(1+1/N)\mathbf{P}$, instead of $\mathbf{P}$, because the ensemble mean does not coincide with the mean of the Gaussian distribution.

3.6 Alternate ETKF-N

The alternative formulation of ETKF-N, that assumes $\overline{\boldsymbol{x}}$ is the best estimator for the prior, leads to the background term

$$\widetilde{\mathcal{J}}_b^{\mathrm{alt}} = \frac{N}{2}\ln\left(1+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right). \tag{46}$$

The only difference is in the missing $1/N$ offset term, which is not surprising since it was identified as a measure of the mistrust in the ensemble mean to represent the true forecast mean.

4 Tests and validation with simple models

In this section, the new filters will be numerically tested, on a three-variable chaotic dynamical toy model, as well as a one-dimensional chaotic dynamical toy model. For the numerical experiments, $\mathbf{U}$ is chosen to be the identity.

4.1 Lorenz '63 toy-model

4.1.1 Setup

The Lorenz '63 model (Lorenz, 1963) is a model with $M = 3$ variables, defined by the equations:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y-x)\\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \rho x - y - xz\\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - \beta z.\end{aligned} \tag{47}$$

The parameters are set to the original values $(\sigma,\rho,\beta) = (10, 28, 8/3)$, which are known to lead to chaotic dynamics, with a doubling time of 0.78 time units. In the following simulations, a reference simulation stands for the truth. The model is considered to be perfect: the model of the truth is the same as the one used in data assimilation runs. We generate synthetic observations from the reference simulation for the three variables each Δt time interval, with $\Delta t = 0.10$, $\Delta t = 0.25$ and $\Delta t = 0.50$. These choices are expected to generate mild, medium and strong impact of non-linearity and, as a possible consequence, non-Gaussianity of errors. These observations are independently perturbed with a normal white noise of standard deviation 2 following Harlim and Hunt (2007). In comparison, the natural variability (standard deviation from the mean of a long model run) of the x, y, and z variables is 7.9, 9.0, and 8.6 respectively.

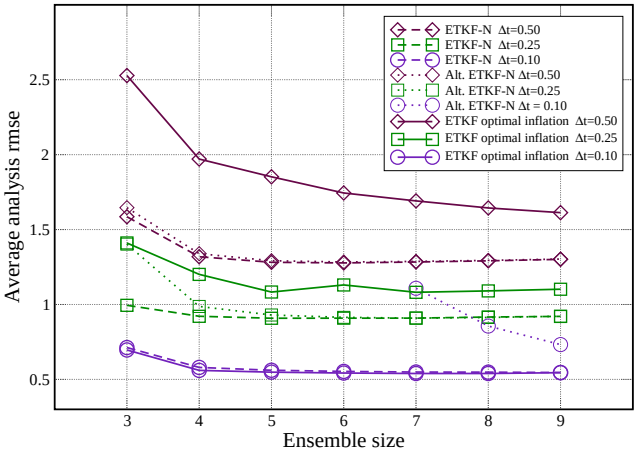

Fig. 1. Time-averaged analysis rmse for ETKF, ETKF-N and the alternate ETKF-N, for three experiments with different time intervals between updates, and for an ensemble size from $N = 3$ to $N = 9$.

All the simulations are run for a period of time corresponding to 5×10^5 cycles, for the three values of Δt. We use a burn-in period of 10^4 cycles to minimize any impact on the final result. The ensemble size is varied from $N = 3$ to $N = 9$. The filters are judged by the time-averaged value of the root mean square error between the analysis and the true state of the reference run.

4.1.2 Best rmse

For ETKF, a multiplicative inflation is applied by rescaling of the ensemble deviations from the mean:

$$\boldsymbol{x}_k \longrightarrow \overline{\boldsymbol{x}} + r\,(\boldsymbol{x}_k - \overline{\boldsymbol{x}}), \tag{48}$$

so that $r = 1$ means no inflation. A wide range of inflation factors r is tested. The inflation factor leading to the smallest (best) rmse is selected. For finite-size filters, inflation is not considered. Therefore for each finite-size filter score, only one run is necessary.

The results are reported in Fig. 1.

For mild non-linearity, the ETKF is slightly better than ETKF-N. With a stronger impact of non-linearity ($\Delta t = 0.25$ and $\Delta t = 0.50$), ETKF-N significantly outperforms ETKF. The alternate ETKF-N is diverging for $\Delta t = 0.10$ and for small ensemble size $N \leq 6$. This emphasizes the fact that the ensemble mean $\overline{\boldsymbol{x}}$ is not a fine estimation of $\boldsymbol{x}_b$, the mean

of the true error distribution. For $\Delta t = 0.25$ and $\Delta t = 0.50$, where the errors are larger and the estimation of $\boldsymbol{x}_b$ may be relatively less important, the performance is almost as good as ETKF-N, with slight deviations for the smallest ensembles.

To a large extent, these results are similar to those of Harlim and Hunt (2007). However, even though getting better results than ETKF, their filter still necessitates to adjust one parameter.

4.2 Lorenz '95 toy-model

4.2.1 Setup

The filters are also applied to the one-dimensional Lorenz '95 toy-model (Lorenz and Emmanuel, 1998). This model represents a mid-latitude zonal circle of the global atmosphere, discretized into $M = 40$ variables $\{x_m\}_{m=1,\ldots,M}$. The model reads, for $m = 1, \ldots, M$,

$$\frac{\mathrm{d}x_m}{\mathrm{d}t} = (x_{m+1} - x_{m-2})x_{m-1} - x_m + F\,, \tag{49}$$

where $F = 8$, and the boundary is cyclic. Its dynamics is chaotic, and its attractor has a topological dimension of 13, a doubling time of about 0.42 time units, and a Kaplan-Yorke dimension of about 27.1.

The experiments follow the configuration of Sakov and Oke (2008). In the first experiment, the time interval between analyzes is $\Delta t = 0.05$, representative of time intervals of 6 hours for a global meteorological model. With this choice, non-linearity mildly affects the dynamics between updates. All variables are observed every Δt. Therefore, the observation operator is the identity matrix. All observations, which are obtained from a reference model run (the truth), are perturbed with a univariate normal white distribution of standard deviation 1. The observation error prior is accordingly a normal distribution of error covariance matrix the identity. In comparison, the natural variability of the model (standard deviation from the mean) is 3.6 for any of the $M = 40$ variables. The performance of a filter is assessed by the root mean square error (rmse) of the analysis with the truth, averaged over the whole experiment run.

As a burn-in period, 5×10^3 analysis cycles are used, whereas 10^4 analysis cycles are used for the assimilation experiments. This may be considered relatively short. However, on the one hand, the convergence was deemed sufficient for this demonstrative study. On the other hand, about 5×10^4 assimilation experiments have been performed, because the inflation (and later the localization) parameters are investigated for many sizes of the ensemble. Longer runs (10^5 analysis cycles) have also been performed, but no (long-term) instability was noted. Moreover these tests showed that the rmses of the 10^4-cycle cases had reasonably converged.

4.2.2 Ensemble size – inflation diagrams

Following Sakov and Oke (2008) and many others, we investigate the rmse of the analysis with the reference state (the truth). The ensemble size is varied from 5 to 50. A multiplicative inflation is applied by rescaling of the ensemble deviations from the mean according to Eq. (48). The inflation factor r is varied from 1. to 1.095 by step of 0.005. As a result, one obtains two-dimensional tables of rmse, which are displayed graphically.

The results for ETKF are reported in Fig. 2a. They are similar to the symmetric ensemble square root Kalman filter of Sakov and Oke (2008). The filter starts converging when the ensemble size is larger than the model unstable subspace dimension. Inflation is always necessary, even for a size of the ensemble greater than the Kaplan-Yorke dimension pointing to a systematic underestimation of sampling errors.

The results of ETKF-N are reported on Fig. 2b. At first, it is striking that the filter diverges for ensemble sizes below $N = 15$. This is disappointing, since the original goal of this study was to remedy to all sampling flaws in a deterministic context. This is obviously not achieved, similarly to any kind of EnKF without localization. However, the formalism developed here allows to understand the reason of this failure, and how it could later be amended. This will be discussed in Sect. 5.

Beyond $N = 15$ (which corresponds to a rank of 14 or less from the anomaly subspace, close to the model unstable subspace dimension), the filter is unconditionally stable.

The results of the alternate ETKF-N are also reported on Fig. 2c. It is also unconditionally stable beyond $N = 15$, but the rmses are better.

4.2.3 Best rmse

In the case of ETKF, the best root mean square error is obtained by taking, for each ensemble size, the minimal rmse over all inflation factors. For ETKF-N, there is only one rmse, since inflation is not considered. In Fig. 3 are plotted the best rmses for the three filters. The alternate ETKF-N seems to outperforms ETKF slightly. But its major asset is that the alternate ETKF-N obtains the best rmses without inflation.

Both ETKF-N and alternate ETKF-N perform better than ETKF over the range $N = 5$–16, especially in the critical range $N = 14$–16. This has been checked for other configurations (changing M and F) of the Lorenz '95 model.

The ETKF-N is not as good as the other two filters beyond $N = 16$. It underperforms both filters by a maximum of 30 %, for $N = 20$. We believe this is explained by the robustness of the filter. Indeed, as discussed in Sect. 3, ETKF-N assumes that the mean state can be different from the mean state of the true distribution, whereas the alternate ETKF-N assumes they match. Both finite-size filters are symmetric. If the model flow remains approximately linear, which is the

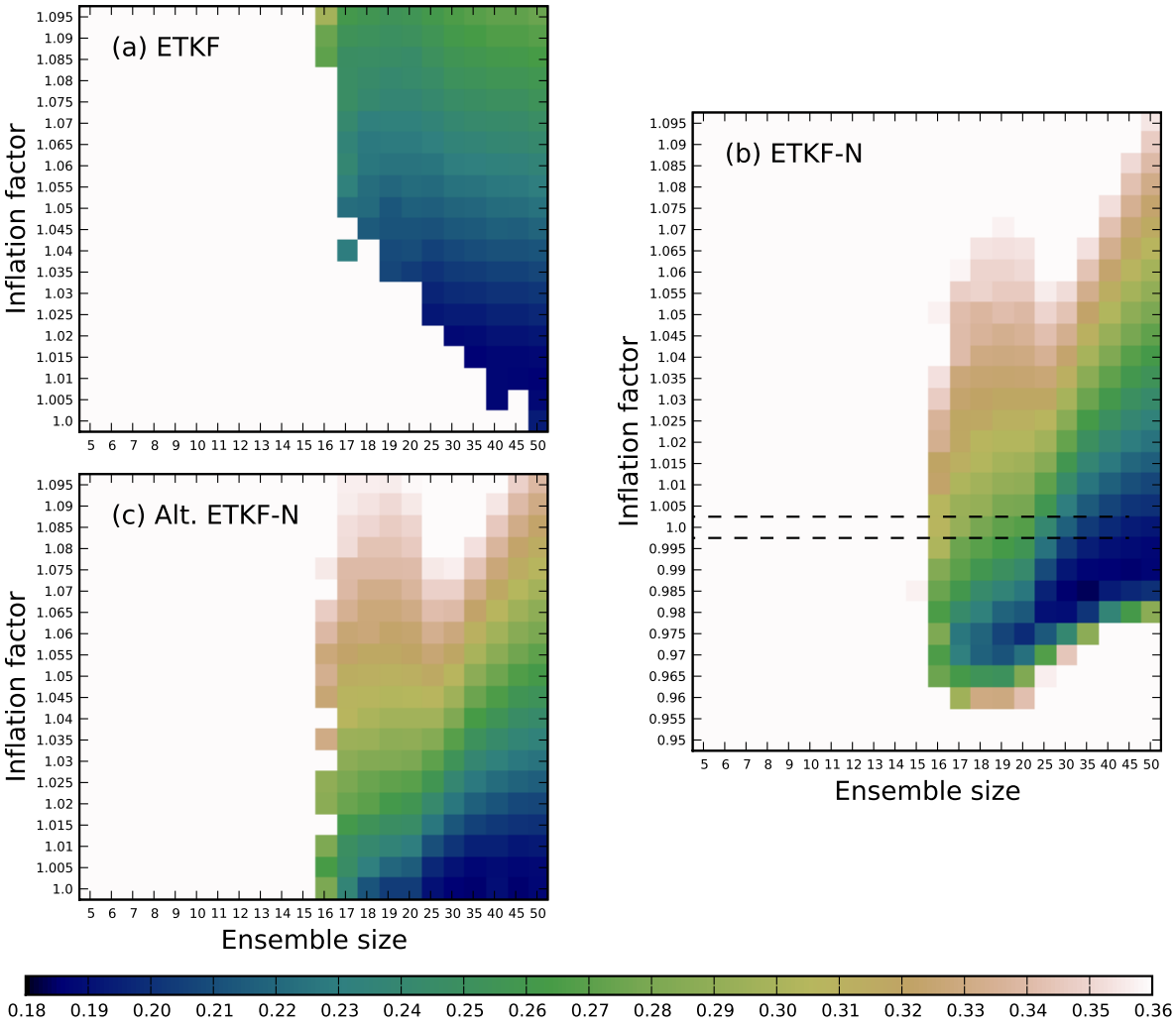

Fig. 2. Root mean square errors of ETKF **(a)**, ETKF-N **(b)**, and alternate ETKF-N **(c)**, for a wide range of ensemble size (5–50) and inflation parameters $r = 1, 1.005, \ldots, 1.095$ and in panel **(b)** a larger range of inflation/deflation $r = 0.945, \ldots, 1.095$.

case for small Δt, the forecasted ensemble will remain centered on the trajectory of the mean, so that the mean of the ensemble will remain a good estimate of the true distribution mean. Therefore, the alternate ETKF-N, as well as symmetric ensemble square root filters, have an advantage in linear conditions over the more conservative, too cautious ETKF-N. If this is correct, then the performance of ETKF-N (which is symmetric) should be better, or at worst equal to the performance of a non-symmetric ensemble square root Kalman filter for small Δt. Indeed this can be checked by comparison of Fig. 3 of the present article and Fig. 4 of Sakov and Oke (2008). Moreover, according to this argument, the performance of ETKF-N should improve for larger ensemble size and larger Δt, in comparison with ETKF (with optimal inflation).

As mentioned earlier, the time interval between updates has been set to $\Delta t = 0.05$. We know from the experiment on the Lorenz '63 model and from the previous remark, that the performance of ETKF-N as compared to ETKF is susceptible to vary with Δt. Let us take the example of an ensemble size of $N = 20$. The setup is unchanged except for the time interval which is set to $\Delta t = 0.05, 0.10, \ldots, 0.30$.

As shown in Fig. 4, $\Delta t \leq 0.15$ is a turning point beyond which ETKF-N obtains better rmse than ETKF without inflation. The alternate ETKF-N offers the best of ETKF (with optimal inflation) and ETKF-N, over the full range of Δt.

5 Local extension of ETKF-N

5.1 Can the use of localization be avoided?

We saw numerical evidence that ETKF-N does not solve the full sampling problem. A similar conclusion can be reached from a more mathematical standpoint. The functional form

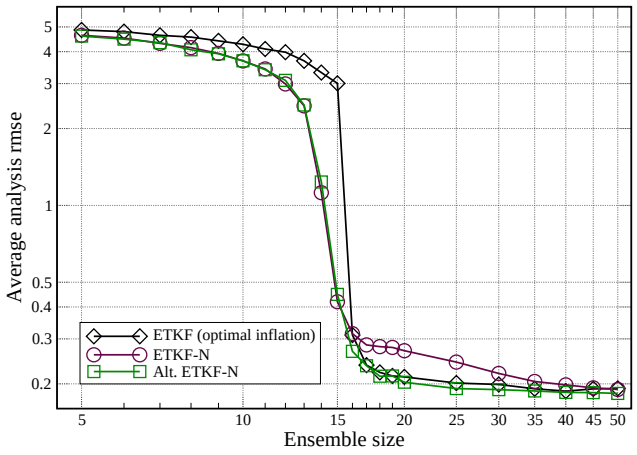

Fig. 3. Best rmse over a wide range of inflation factors for ETKF, and rmse without inflation for ETKF-N and for the alternate ETKF-N.

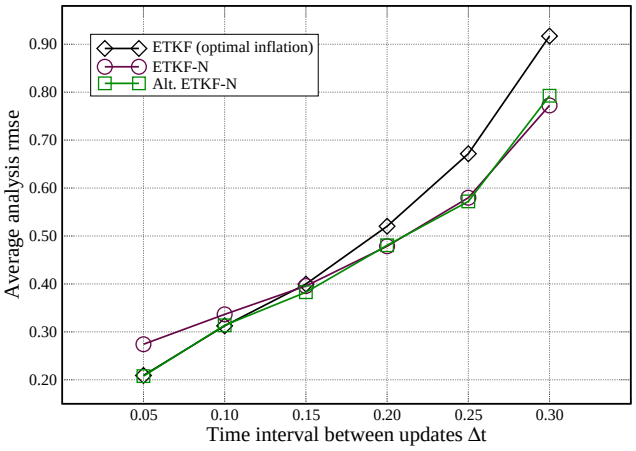

Fig. 4. Best rmse for ETKF over a wide range of inflation factors, rmse of ETKF-N without inflation and rmse of the alternate ETKF-N without inflation, for several time intervals between updates, and for an ensemble size $N = 20$.

of ETKF-N Eq. (35) is formulated in ensemble space, via a set of ensemble coordinates that do not depend on the real space position. This functional form is due to the choice of the Jeffrey's prior. It implies that the dimension of the analysis space has a very reduced rank. Localization, which is meant to increase this rank, is therefore mandatory below some context-dependent ensemble size. We could contemplate two ways to tackle this difficult problem.

The first one would consist in trading Jeffrey's prior for a more informative one. The particular form of the ETKF-N background term which depends only on the ensemble coordinates was due to the specific choice of Jeffrey's prior, which had the merit to be simple. However, errors of the Lorenz '95 data assimilation system, or of more realistic geophysical systems, often have short-range correlations. If, using an hyperprior different from Jeffreys', one could integrate on a restricted set of error covariance matrices of correlation matching the climatological correlations of the data assimilation system, we conjecture that localization could be consistently achieved within the proposed formalism.

Let us take an example where it is assumed that the correlations of the data assimilation system are very short-range. At the extreme, we integrate Eq. (14) on the set of all positive definite diagonal matrices $\mathbf{B}$, that is a set of M positive scalars, with the non-informative univariate prior:

$$p_J([\mathbf{B}]_{ii}) = [\mathbf{B}]_{ii}^{-1}. \tag{50}$$

Following the derivation of Sect. 2, one obtains

$$\mathcal{J}_b(\boldsymbol{x}) = \frac{N}{2}\sum_{i=1}^{M}\ln\left[\frac{N}{N+1}(x_i-\overline{x}_i)^2+\sum_{k=1}^{N}(x_i^k-\overline{x}_i)^2\right]. \tag{51}$$

As opposed to the background terms introduced earlier, this $\mathcal{J}_b$ cannot be written in ensemble space, i.e. not in terms of the coordinates w_k in the vector space of anomalies. Assuming the same setup used for the Lorenz '95 model, the average analysis rmse of such EnKF-N is in the range 0.50 for $N = 5$ down to 0.35 for $N = 40$. It has been checked to be similar to any EnKF or ETKF with a minimal (meaningful) localization length, except that, for this new filter, inflation is not necessary even for small N. We conclude that localization can potentially be expressed in the formalism. Pursuing this idea is well beyond the scope of this article, because it seems mathematically challenging.

As an alternative, simpler, and widespread solution, a local version of the filter will be developed and tested in the remaining of this section.

5.2 LETKF-N

The extension of ETKF-N to a local ensemble transform Kalman filter is the same as the passage from ETKF to LETKF as described by Hunt et al. (2007), and by Harlim and Hunt (2007) for non-quadratic cost functions. For high-dimensional and computationally challenging systems, this requires to follow their algorithm. However, for the Lorenz '95 toy-model the passage from ETKF-N to LETKF-N is trivial. Indeed, fixing a localization length l for each point of control space, one performs a local analysis using all observations within a radius of l units. Hence, for the Lorenz '95 model, l ranges from $l = 0$, using the single local observation if any, to $l = 20$, meaning that all observations are assimilated, i.e. no localization. We shall call this filter the finite-size (finite-sample) local ensemble transform Kalman filter (LETKF-N).

5.3 Application to the Lorenz '95 toy-model

5.3.1 Ensemble size – inflation diagrams

The results of the experiments with LETKF, LETKF-N, and with the alternate LETKF-N are reported in Fig. 5.

For LETKF, inflation is still necessary to stabilize the filter for not-so-small ensemble sizes ($N \leq 11$). LETKF-N does not require inflation (at least for $N \geq 5$, the case $N < 4$ was not investigated). Again, it means that LETKF-N estimates well, or over-estimates, sampling errors. But it is unconditionally stable with the best performance obtained without inflation. The alternate LETKF-N may still be the best filter for an ensemble size beyond the model unstable subspace dimension, but it disappoints by requiring inflation for small and moderate ensemble size. This indicates that trusting the mean $\overline{\boldsymbol{x}}$ as the first guess is a source of error for small ensemble size.

5.3.2 Best rmse

In Fig. 6 are plotted the best rmses for the three filters, with localization.

LETKF-N is always slightly suboptimal (with a maximal discrepancy of 10 % for $N = 5$). However, it is the only unconditionally stable filter of the three: it does not require inflation.

The alternate LETKF-N is as good as LETKF-N for small ensembles but it does require inflation, which is why it is not so interesting in this regime. The alternate LETKF-N is as good as LETKF in the *large* ensemble size regime, but without inflation.

As we increase the time interval between updates, the performance of the filters degrades but their relative performance evolves. Let us take the example of an ensemble size of $N = 10$, following Harlim and Hunt (2007). The setup is unchanged except for the time interval between updates which is set to $\Delta t = 0.05, 0.10, \ldots, 0.50$. The results are reported in Fig. 7.

For $\Delta t \leq 0.20$, LETKF with optimal inflation and localization outperforms LETKF-N with optimal localization and no inflation. For $\Delta t \geq 0.20$, LETKF-N dominates. Like in the Lorenz '63 case, this is reminiscent of the results of Harlim and Hunt (2007). This indicates that the relative performance of filters as shown for instance by Fig. 6 should not be taken as a rule, since there are regimes where LETKF-N (without inflation) performs better than LETKF.

The good performances of EnKF-Ns relative to the EnKFs in the strong nonlinear regime, is not an indication that EnKF-N can handle non-Gaussianity in this regime. However the sampling errors may be created and exacerbated by the non-linearity of the model flow, and hence of the non-Gaussianity of the underlying pdf of errors. This may give an advantage to the finite-size ensemble filters in this regime.

6 Summary and future directions

Current strategies for stabilizing the ensemble Kalman filter are empirical tuning of inflation, use of multi-ensemble configuration, explicit identification of the sampling/model errors, or adaptive optimization of inflation. In this article, we have followed a somehow different route. A new background prior pdf that takes into account the discrete nature of the ensemble was derived using Bayesian principles. It accounts for the uncertainty attached to the first- and second-order empirical moments of the ensemble seen as a sample of a true error distribution. The definition of the prior pdf leads to a new class of filters (EnKF-N). Even though the resulting prior is non-Gaussian, it is entirely based on Gaussian hypotheses for the errors. In principle, through this prior, the analysis should take into account sampling errors.

Specifically, an ensemble transform variant (ETKF-N) was derived in the spirit of the ETKF of Hunt et al. (2007). It is tested on the Lorenz '63 and the Lorenz '95 toy models. In the absence of model error, the filter appear to be stable without inflation for ensemble size greater than the model unstable subspace dimension of the attractor. Moreover, for large enough time interval between updates, its performance is superior to that of ETKF. A variant of ETKF-N is expected to outperform ETKF-N for small time interval between updates: without inflation, it seems to systematically perform as well as, or better than ETKF. Unfortunately, as shown in the case of the Lorenz '95 model, these finite-size filters diverge for smaller ensemble size, like any ensemble Kalman filter. Localization is mandatory. That is why a local variant of the filter (LETKF-N) which parallels LETKF, is developed.

From experiments on the Lorenz '95 toy model, LETKF-N scheme seems stable without inflation. Depending on the time interval between updates, its performance with optimally tuned localization can be slightly inferior or superior to LETKF with optimally tuned localization and optimally tuned inflation.

The methodology presented here is mainly a *proof of concept*. We believe that more work is needed to explore the strengths and limitations of the methodology, and that there is room for improvement of the schemes.

For instance, we conjectured that the incapacity of ETKF-N to fully account for sampling errors (as opposed to LETKF-N with optimally tuned localization), was mainly due to the use of an hyperprior which generates correlations different from that of the data assimilation system built on the particular model. To avoid using weakly informative (hyper)priors on $\boldsymbol{x}_b$ and $\mathbf{B}$, one solution is to pass information between analyzes beyond the knowledge of the ensemble. In the context of oil reservoir monitoring, Myrseth and Omre (2010) have built a sophisticated and elegant ensemble Kalman filter that could be seen as a stochastic extension of ETKF-N that achieves such a goal. They see covariance matrices as random matrices with an inverse Wishart distribution of precision matrix Ψ in $\mathbb{R}^{M\times M}$ (Muirhead, 1982).

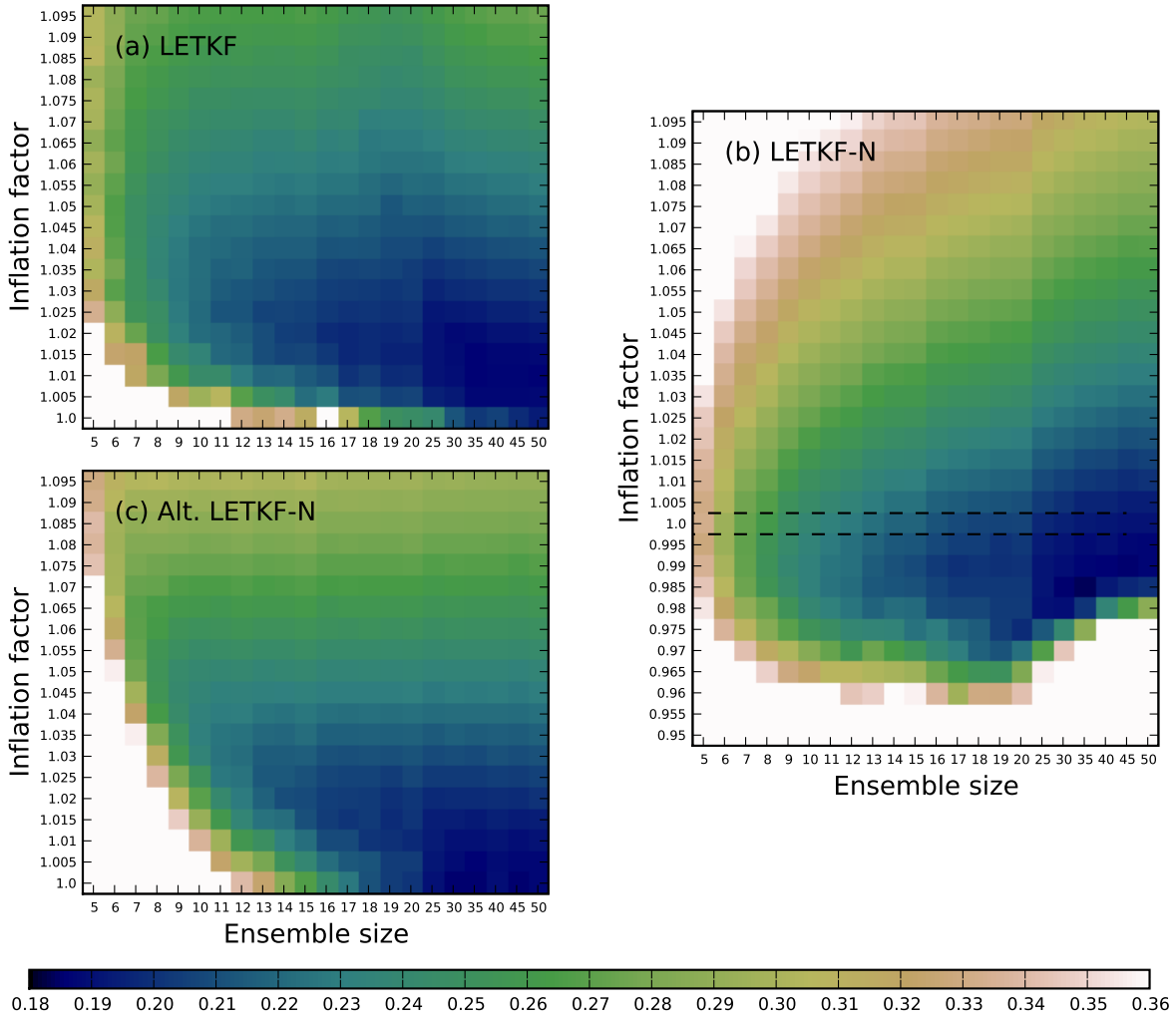

Fig. 5. Root mean square errors of LETKF **(a)**, LETKF-N **(b)**, and alternate LETKF-N **(c)**, for a wide range of ensemble size (5–50) and inflation factors $r = 1, 1.005, \ldots, 1.095$ and in panel **(b)** a larger range of inflation/deflation $r = 0.945, \ldots, 1.095$.

The hyperprior for $\boldsymbol{x}_b$ and $\mathbf{B}$ are chosen in such a way (conjugate distribution) that the posterior error covariance matrix still follows an inverse-Wishart distribution. However, such a $\mathbf{B}$ matrix should be drawn from this distribution for each member, and the corresponding innovation statistics computed and inverted, which could become very costly. Even though one assumes all members use the same draw of $\mathbf{B}$, it is necessary to store the precision matrix Ψ, which cannot be afforded in the high-dimensional context of geophysics. Still, to pass supplementary information (beyond the ensemble) one might contemplate adapting the algorithm of Myrseth and Omre (2010) so as to maintain a reduced-order, short memory, precision matrix, with a rank of a few ensemble sizes.

The behavior of EnKF-N in limiting regimes is worth exploring. For instance, in the limiting case where the dynamical model is linear, EnKF-N may not exhibit optimal performance since EnKF-N does not make implicit assumptions on the linearity of the model as opposed to traditional EnKFs. In this limiting case, the hyperprior $p(\boldsymbol{x}_b, \mathbf{B})$ should optimally be a Dirac delta function pdf peaked at the empirical moments of the ensemble, which would make EnKF-N the traditional EnKF. But what happens to EnKF-N with Jeffreys' hyperprior in this regime is less clear.

Another lead for improvement points to the derivation of the new prior used in the (L)ETKF-N filters, which, by definition, ignores the observations to be assimilated. From a Bayesian perspective, this is suboptimal: in our scheme, any $\mathbf{B}$ matrix's likelihood is tested against the ensemble, but not against both the ensemble and the observations, which is what a full Bayesian scheme would prescribe. Indeed Eq. (6) should be generalized to:

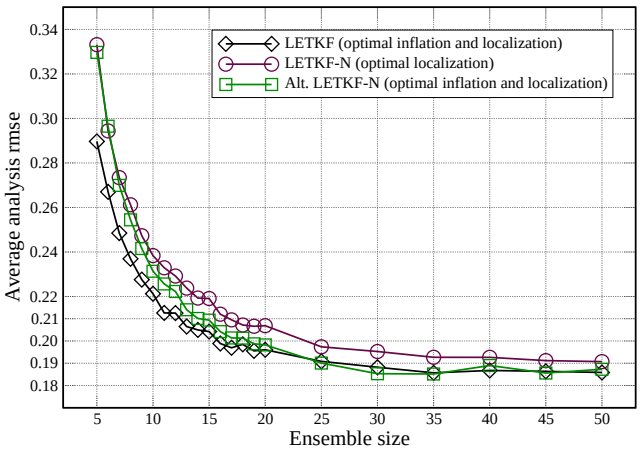

Fig. 6. Best rmse over a wide range of inflation factors, and all possible localization lengths for LETKF and the alternate LETKF-N. Best rmse without inflation over all possible localization lengths for LETKF-N.

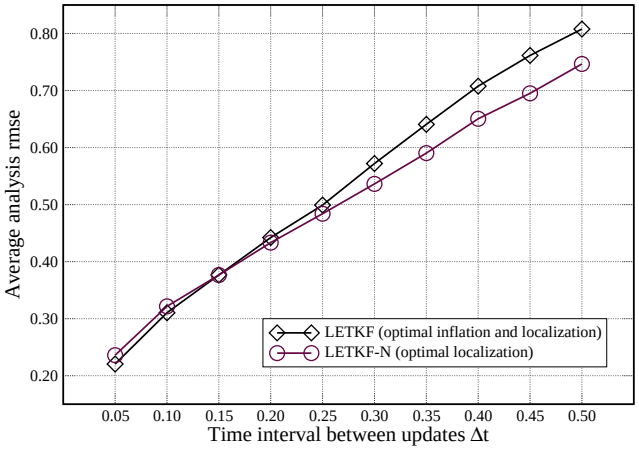

Fig. 7. Best rmse for LETKF and LETKF-N over all possible localization lengths and a wide range of inflation factors for LETKF, for several time intervals between updates, and for an ensemble size $N = 10$.

$$
\begin{aligned}
&p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\\
&=\int \mathrm{d}\boldsymbol{x}_b\mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{x}_b,\mathbf{B})\,p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{y},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\\
&=\int \mathrm{d}\boldsymbol{x}_b\mathrm{d}\mathbf{B}\, p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{x}_b,\mathbf{B})\,p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\\
&=p(\boldsymbol{y}|\boldsymbol{x})\int \mathrm{d}\boldsymbol{x}_b\mathrm{d}\mathbf{B}\,\frac{p(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})}{p(\boldsymbol{y}|\boldsymbol{x}_b,\mathbf{B})}\,p(\boldsymbol{x}_b,\mathbf{B}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N). \qquad (52)
\end{aligned}
$$

Because of the presence of $p(\boldsymbol{y}|\boldsymbol{x}_b,\mathbf{B})$ in the last integral and its dependence in $\boldsymbol{x}_b$ and $\mathbf{B}$, it seems difficult to analytically solve the problem in order to generalize ETKF-N.

Stability without inflation is a property shared by ETKF-N, or LETKF-N, with the multi-ensemble configuration of Mitchell and Houtekamer (2009). However the two approaches draw their rationale from two different standpoints: Bayesian statistics and cross-validation respectively, whose connections are not clearly understood in Statistics. Additionally, we note that the multi-ensemble approach makes use of the observation while the finite-size ensemble transform filters do not. In other words, the multi-ensemble approach reduces the errors in the analysis while the finite-size ensemble transform filters do so prior to the analysis. The methodology developed in this article naturally led to deterministic filters, whose comparison with stochastic filters cannot be simple. Therefore it would be interesting to develop a stochastic filter counterpart to the deterministic EnKF-N presented here.

The focus of this study was primarily on sampling errors. In a realistic context, one should additionally take into account model errors, and the errors that come from the deviation from Gaussianity due to model non-linearity. If one trusts from the previous results that EnKF-N reduces significantly the need for inflation meant to compensate for sampling errors, then the use of inflation in EnKF-N would essentially be a measure of model errors. It could also be a measure of the deviation from Gaussianity, or of the misspecification of the hyperprior as discussed earlier. These ideas have been successfully tested on the context of the Lorenz '95 using the setup of this article. However, reporting these results is beyond the scope of this article.

As a final remark, we would like to mention that the prior pdf $p(\boldsymbol{x}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \propto \exp(-\mathcal{J}_b(\boldsymbol{x}))$, where $\mathcal{J}_b$ is defined by Eq. (21), could more generally be useful in environmental statistical studies, when one needs to derive a pdf from samples of the system state, or of some error about it, which is assumed Gaussian-distributed. Note that the ensemble size needs to be large enough otherwise localization is still necessary.

Appendix A

One minimum or more

Here, the observation operator is supposed to be linear or linearized. Define the innovation $\boldsymbol{\delta} = \boldsymbol{y} - \mathbf{H}\overline{\boldsymbol{x}}$. Equation (35) can be written

$$
\begin{aligned}
\widetilde{\mathcal{J}}_a(\boldsymbol{w}) = &\frac{1}{2}(\boldsymbol{\delta}-\mathbf{H}\mathbf{X}\boldsymbol{w})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\delta}-\mathbf{H}\mathbf{X}\boldsymbol{w})\\
&+\frac{N}{2}\ln\left(1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right). \qquad (A1)
\end{aligned}
$$

If there is a minimum, it must satisfy

$$
\left[(\mathbf{H}\mathbf{X})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{X}+\frac{N}{1+\frac{1}{N}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}\right]\boldsymbol{w}=(\mathbf{H}\mathbf{X})^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}. \qquad (A2)
$$

In order to diagonalize the left-hand side matrix, we can use the singular value decomposition

$$(\mathbf{HX})^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \tag{A3}$$

where $\mathbf{U}$ is an orthogonal matrix in $\mathbb{R}^{N\times N}$, $\mathbf{V}$ is in $\mathbb{R}^{d\times N}$ and satisfies $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_N$, and $\mathbf{D}$ is a diagonal matrix in $\mathbb{R}^{N\times N}$. Define $\widehat{\boldsymbol{w}} = \mathbf{U}^{\mathrm{T}}\boldsymbol{w}$, then

$$\left[\mathbf{D}^2 + \frac{N}{1+\frac{1}{N}+\widehat{\boldsymbol{w}}^{\mathrm{T}}\widehat{\boldsymbol{w}}}\right]\widehat{\boldsymbol{w}} = \mathbf{D}\boldsymbol{v}, \tag{A4}$$

where $\boldsymbol{v} = \mathbf{V}^{\mathrm{T}}\mathbf{R}^{-1/2}\boldsymbol{\delta}$. Then the left-hand side matrix is diagonal.

In the case of a single observation (serial assimilation), $\mathbf{D}$ has only one non-zero entry (call it α), and $\mathbf{D}\boldsymbol{v}$ is a vector with at most one non-zero entry (call it β). Then solving for all the other components, it is clear that the components of $\widehat{\boldsymbol{w}}$ not related to α are zero. Then the remaining scalar equation for the non-trivial component γ of $\widehat{\boldsymbol{w}}$ is

$$\left[\alpha^2 + \frac{N}{1+\frac{1}{N}+\gamma^2}\right]\gamma = \beta. \tag{A5}$$

This third-order algebraic equation in γ has either one real solution or three real solutions. Therefore, the cost function $\widetilde{\mathcal{J}}_a$ has a global minimum, and possibly another local minimum. Note that with several observations assimilated in parallel, there may be more local minima.

Acknowledgements. The author acknowledges a useful discussion with C. Snyder on state-of-the-art ensemble Kalman filtering. The article has benefited from the useful comments and suggestions of L. Delle Monache, N. Bowler, P. Sakov acting as a Reviewer, an anonymous Reviewer, O. Talagrand acting as Editor, and L. Wu.

Edited by: O. Talagrand
Reviewed by: P. Sakov and another anonymous referee

References

Anderson, J. L.: An ensemble adjustment Kalman filter for data assimilation, Mon. Weather Rev., 129, 2884–2903, 2001.

Anderson, J. L.: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, Physica D, 230, 99–111, 2007a.

Anderson, J. L.: An adaptive covariance inflation error correction algorithm for ensemble filters, Tellus A, 59, 210–224, 2007b.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, Mon. Weather Rev., 127, 2741–2758, 1999.

Bishop, C. H. and Hodyss, D.: Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation, Q. J. Roy. Meteor. Soc., 133, 2029–2044, 2007.

Bocquet, M., Pires, C. A., and Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation, Mon. Weather Rev., 138, 2997–3023, 2010.

Brankart, J.-M., Cosme, E., Testut, C.-E., Brasseur, P., and Verron, J.: Efficient adaptive error parameterization for square root or ensemble Kalman filters: application to the control of ocean mesoscale signals, Mon. Weather Rev., 138, 932–950, 2010.

Burgers, G., van Leeuwen, P. J., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Mon. Weather Rev., 126, 1719–1724, 1998.

Carrassi, A., Vannitsem, S., Zupanski, D., and Zupanski, M.: The maximum likelihood ensemble filter performances in chaotic systems, Tellus A, 61, 587–600, 2009.

Corazza, M., Kalnay, E., and Patil, D.: Use of the breeding technique to estimate the shape of the analysis errors of the day, J. Geophys. Res., 10, 233–243, 2002.

Dee, D. P.: On-line Estimation of Error Covariance Parameters for Atmospheric Data Assimilation, Mon. Weather Rev., 123, 1128–1145, 1995.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, Q. J. Roy. Meteor. Soc., 131, 3385–3396, 2005.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10143–10162, 1994.

Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, Springer-Verlag, 2nd Edn., 2009.

Furrer, R. and Bengtsson, T.: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivariate Anal., 98, 227–255, 2007.

Gejadze, I. Y., Le Dimet, F.-X., and Shutyaev, V.: On analysis error covariances in variational data assimilation, SIAM J. Sci. Comput., 30, 1847–1874, 2008.

Gottwald, G. A. and Mitchell, L., and Reich, S.: Controlling overestimation of error covariance in ensemble Kalman filters with sparse observations: A variance-limiting Kalman filter, Mon. Weather Rev., 139, 2650–2667, 2011.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, Mon. Weather Rev., 129, 2776–2790, 2001.

Harlim, J. and Hunt, B.: A non-Gaussian ensemble filter for assimilating infrequent noisy observations, Tellus A, 59, 225–237, 2007.

Houtekamer, P. L. and Mitchell, H. L.: Data assimilation using an ensemble Kalman filter technique, Mon. Weather Rev., 126, 796–811, 1998.

Houtekamer, P. L. and Mitchell, H. L.: A sequential ensemble Kalman filter for atmospheric data assimilation, Mon. Weather Rev., 129, 123–137, 2001.

Huber, P. J.: Robust regression: Asymptotics, conjectures, and Monte Carlo, Ann. Statist., 1, 799–821, 1973.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112–126, 2007.

Jeffreys, H.: Theory of Probability, Oxford University Press, 3rd Edn., 1961.

Lei, J., Bickel, P., and Snyder, C.: Comparison of Ensemble Kalman Filters under Non-Gaussianity, Mon. Weather Rev., 138, 1293–1306, 2010.

Li, H., Kalnay, E., and Miyoshi, T.: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, Q. J. Roy. Meteor. Soc., 135, 523–533, 2009.

Livings, D. M., Dance, S. L., and Nichols, N. K.: Unbiased ensemble square root filters, Physica D, 237, 1021–1028, 2008.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, 1963.

Lorenz, E. N. and Emmanuel, K. E.: Optimal sites for supplementary weather observations: simulation with a small model, J. Atmos. Sci., 55, 399–414, 1998.

Mallat, S., Papanicolaou, G., and Zhang, Z.: Adaptive covariance estimation of locally stationary processes, Ann. Stat., 26, 1–47, 1998.

Mitchell, H. L. and Houtekamer, P. L.: An adaptive ensemble Kalman filter, Mon. Weather Rev., 128, 416–433, 1999.

Mitchell, H. L. and Houtekamer, P. L.: Ensemble Kalman Filter Configurations and Their Performance with the Logistic Map, Mon. Weather Rev., 137, 4325–4343, 2009.

Muirhead, R. J.: Aspect of Multivariate Statistical Theory, Wiley-Interscience, 1982.

Myrseth, I. and Omre, H.: Hierarchical Ensemble Kalman Filter, SPE J., 15, 569–580, 2010.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D. J., and Yorke, A.: A local ensemble Kalman filter for atmospheric data assimilation, Tellus A, 56, 415–428, 2004.

Raynaud, L., Berre, L., and Desroziers, G.: Objective filtering of ensemble-based background-error variances, Q. J. Roy. Meteor. Soc., 135, 1177–1199, 2009.

Sacher, W. and Bartello, P.: Sampling Errors in Ensemble Kalman Filtering. Part I: Theory, Mon. Weather Rev., 136, 3035–3049, 2008.

Sakov, P. and Bertino, L.: Relation between two common localisation methods for the EnKF, Comput. Geosci., 15, 225–237, 2010.

Sakov, P. and Oke, P. R.: Implications of the Form of the Ensemble Transformation in the Ensemble Square Root Filters, Mon. Weather Rev., 136, 1042–1053, 2008.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S.: Ensemble square root filters, Mon. Weather Rev., 131, 1485–1490, 2003.

van Leeuwen, P. J.: Comment on Data Assimilation Using an Ensemble Kalman Filter Technique, Mon. Weather Rev., 127, 1374–1377, 1999.

Wang, X., Bishop, C. H., and Julier, S. J.: Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble?, Mon. Weather Rev., 132, 1590–1605, 2004.

Whitaker, J. S. and Hamill, T. M.: Ensemble Data Assimilation without Perturbed Observations, Mon. Weather Rev., 130, 1913–1924, 2002.

Wick, G. C.: The Evaluation of the Collision Matrix, Phys. Rev., 80, 268–272, 1950.

Zhou, Y., McLaughlin, A., and Entekhabi, D.: Assessing the performance of the ensemble kalman filter for land surface data assimilation, Mon. Weather Rev., 134, 2128–2142, 2006.

Zupanski, M.: Maximum Likelihood Ensemble Filter: Theoretical Aspects, Mon. Weather Rev., 133, 1710–1726, 2005.